Empirical Modeling of Nanoscale Dynamics using Solution Mapping

Grant # FA9550-07-0161

Martha Grover
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

February 27, 2010

Abstract

Computer simulations provide useful predictions of complex system dynamics, but they cannot be easily inverted for use in control and optimization. When the computational time to run a single prediction is high, approximate models with reduced computation are required. These models must be straightforward to build and have quantified bounds on their accuracy. With support from this grant two automated methods were developed for building empirical models from simulation data. These methods were subsequently applied to stochastic simulations of nanoscale dynamics.

In the first method, the simulation dynamics are modeled on a discrete state space, with input-dependent transitions between the states. This approach was used for dynamic optimization of a gallium arsenide surface deposition process, which was not computationally feasible for the full simulation. The second method, based on Gaussian process modeling, was developed to further improve the prediction accuracy of the first method, which was limited by the discrete state space. Moreover, Gaussian process modeling enabled a quantification of the prediction variance, which is necessary so that the dynamic model can be used with confidence in control applications.

1 Executive Summary

Complex dynamic simulations are becoming an increasingly central tool in the engineering of advanced technology, and this is just as true in dynamics and control as it is in other areas such as optimization and design. Predictions from simulations can aid in understanding the emergent behavior of a collection of many subsystems, and can be used to validate model predictions and engineering designs. This emergent behavior may stem from a collection of atoms, air vehicles, or human beings.

Simulation alone is not sufficient for the guaranteed robust optimal performance of engineered systems. If simulations for complex systems are to be used in the context of control and/or optimization, there must be a mathematically rigorous framework in which to use these models and simulations. The performance properties and limitations must be proven, just as they are proven for linear systems or for certain low-order deterministic nonlinear systems.

The focus of this specific grant is on stochastic simulations consisting of many-body interactions in nanoscale dynamics, with applications including high-power high-speed transistors and nanoparticle catalysts for energy conversion. Because the dynamics are stochastic, it is not possible to rule

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
28-02-2010	Final Report	2/15/2007 - 11/30/2009

4. TITLE AND SUBTITLE
Empirical Modeling of Nanoscale Dynamics Using Solution Mapping

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-07-0161

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Martha A. Grover

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology
311 Ferst Dr., Atlanta, GA 30332-0100

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
Mathematics, Information and Life Sciences Directorate (NL), Dynamics and Control Program
Program Manager: Dr. Fariba Fahroo
875 N. Randolph St., Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR/NL

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Public availability

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Computer simulations provide useful predictions of complex system dynamics, but they cannot be easily inverted for use in control and optimization. When the computational time to run a single prediction is high, approximate models with reduced computation are required. These models must be straightforward to build and have quantified bounds on their accuracy. With support from this grant two automated methods were developed for building empirical models from simulation data. These methods were subsequently applied to stochastic simulations of nanoscale dynamics. In the first method, the simulation dynamics are modeled on a discrete state space, with input-dependent transitions between the states. This approach was used for dynamic optimization of a gallium arsenide surface deposition process, which was not computationally feasible for the full simulation. The second method, based on Gaussian process modeling, was developed to further improve the prediction accuracy of the first method, which was limited by the discrete state space. Moreover, Gaussian process modeling enabled a quantification of the prediction variance, which is necessary so that the dynamic model can be used with confidence in control applications.

15. SUBJECT TERMS
nanoscale, dynamics, empirical modeling, Gaussian processing modeling

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Martha Grover
U	U	U	SAR	15	19b. TELEPHONE NUMBER *(Include area code)* (404) 894-2878

Reset

anything out for certain, and one can only use statistical measures to quantify performance. This is a departure from much of control theory, in which uncertainties are bounded inside a ball of radius δ and performance is bounded inside a ball of radius ϵ. This assumption of bounded uncertainty is not applicable for many realistic sources of uncertainty—a more typical quantification of uncertainty is the normal (or Gaussian) distribution, in which any outcome is possible, but certain outcomes are more probable. Confidence intervals can then be used to state that an outcome should fall within a specified range with a certain probability.

The overarching goal of this grant is to develop approximate statistical models that reproduce the dynamic evolution in many-body simulations of atomic interactions. Models with reduced computation are needed for use in control and optimization, since a single stochastic realization of the many-body system may take days, weeks, or even longer on a high performance computer. If the underlying models are to be used for control and optimization, the computational time must be reduced dramatically. However, the model accuracy must not be sacrified, at least not beyond the engineering accuracy that is required. Thus, in this research the investigators have developed and implemented a stastical modeling framework that not only predicts the expected mean behavior from the atomic-scale simulation, but also the uncertainty in the prediction. This uncertainty prediction is a key unique feature of the modeling approach, and is required if the approximate models are to be used to engineering tasks like optimization and control.

The basic dynamic equation for the empirical model is

$$\begin{aligned} x[k+1] &= F(x[k], u[k]) \\ t &= k\Delta t \end{aligned} \tag{1}$$

which is sometimes referred to as solution mapping. The goal is then to find the function F, and further to estimate the uncertainty in the prediction $x[k+1]$.

One Ph.D. thesis was completed during the duration of this grant. Cihan Oguz graduated in December 2007, with the thesis entitled "Control-oriented modeling of discrete configuration molecular scale processes: Applications in polymer synthesis and thin film growth." This work was begun under a previous AFOSR grant (FA9550-04-1-0183) and continued with support by the current grant (FA9550-07-0161). The specific outcomes of this thesis include

1. An automated procedure for generating an approximate dynamic model using stochastic simulation data, consisting of the following steps:

   - The high dimensional output of an atomic-scale simulation contains many symmetries and correlations. The independent degrees of freedom in this high dimensional output can be identified and quantified using a pair distribution function (analogous to a Fourier transform) coupled with principal component analysis.
   - A discretization of the state is performed on the subspace of reduced dimension, using clustering techniques.
   - A statistical model is constructed between the discrete states, by defining transitions between the states. These transitions may depend upon the control inputs, and are stored in a matrix, yielding the model structure:

     $$x[k+1] = A(u)x[k] \tag{2}$$

   - The primary computational requirements are associated with the simulations used to identify the model. The resulting model is a linear discrete-time model, enabing virtually instantaneous evaluation. Thus, real-time application in control is practical.

2. Demonstration of this modeling approach in a particular example: a lattice Monte Carlo simulation of ultra-high vacuum deposition of gallium arsenide, a key component of high-speed transistors for communications.

    - The number of independent degrees of freedom observed in the simulation output may be small, indicating an underlying low-dimensional manifold on which the dynamics evolve. In the gallium arsenide example, five principal components were required to accurately reconstruct and then reproduce the dynamics.
    - Because the simulated trajectories lie in a nonlinear submanifold of the five-dimensional space, less than 200 discrete states were sufficient to cover and predict the dynamic evolution of the deposition process. A self-organizing map was used to perform the clustering.
    - A model constructed according to this automated procedure was used for dynamic optimization of the process recipe. The time-varying gallium flux was used as the process input, and the minimum time trajectory to reach a desired surface structure was compute via a dynamic optimization. This optimization would have been infeasible using the full stochastic simulation. However, the optimal trajectory from the reduced model agreed well with stochastic realizations from the full gallium arsenide simulation.

A primary limitation on the accuracy of the approximate model was the discretization of the continuous space. As the number of discrete states is increased, the accuracy improves, but the time required to build the reduced model scales linearly with the number of discrete states, ultimately limiting the number of states that can be included in the approximate model. To further improve the prediction accuracy, a different approach was proposed: conceptually, the idea is to *interpolate among* the discrete set of states in the Markov-chain model.

A second limitation of the method by Oguz is the lack of quantification for the prediction uncertainty in the approximate model. The average error was evaluated and quantified over many typical trajectories, by comparing predictions between the approximate model and the full atomic-scale simulation (having the same initial condition and control input trajectory). However, the uncertainty in the prediction should depend upon the initial state, the control input, and the time.

To address these two limitations, a statistical modeling framework based on Gaussian process modeling (GPM), also known as "kriging," was proposed and implemented under this grant. This approach is the thesis topic of PhD student Andres Hernandez. The key idea in GPM is a two-level modeling approach. First, a deterministic model is selected, which may have unknown parameters that are fit to the stochastic simulation output. This is the typical regression problem. However, in a GPM, a second modeling term is added, which models the *residual* between the simulation output and the first model as a spatially correlated term. This is logical, since the residual between two smooth models should be spatially correlated, and not independent and white, as in the underlying assumption of least-squares fitting. The resulting model will interpolate the sampled data points when the original samples are assumed to be noise-free; otherwise the model prediction *interpolates* the expected mean output. In either case, the GPM also enables computation of the *prediction variance*. Both features are extremely important for black-box empirical modeling.

Gaussian process modeling is a demonstrated and powerful modeling technique in geostatistics, and more recently has been applied extensively in engineering design. However, it is not been exploited in dynamical systems modeling. Recent studies led by Rasmussen [1] show the potential for GPM in dynamics systems modeling, but only for deterministic systems of dimension two. The present work demonstrates the power of GPM in reduced-order modeling, via application in nanoparticle synthesis dynamics.

2 Personnel and Publications

Three graduate students were supported throughout the grant period:

- Cihan Oguz (PhD 2007) was supported during the first year of the grant. His thesis, entitled "Control-oriented modeling of discrete configuration molecular scale processes: applications in polymer synthesis and thin film growth," was the culmination of support from this grant and from the earlier AFOSR grant FA9550-04-1-0183 (2004–2007) from the Dynamics and Control program. This thesis is available electronically from the Georgia Tech library system: http://smartech.gatech.edu/handle/1853/19867.

- Jonathan Rawlston (PhD defense March 2010) was supported part time to perform the stochastic simulation aspects in this grant. His thesis is entitled "Multiscale modeling of free-radical polymerization kinetics."

- Andres Hernandez was supported by this grant for two years, until the end of the grant in November 2009. His PhD thesis topic is kriging for empirical modeling of nanoscale stochastic dynamics.

The research supported by the grant led to the following 5 journal publications and 2 peer-reviewed conference papers:

- C. Oguz and M. A. Gallivan, "Optimization of a thin film process using a dynamic model extracted from molecular simulations," *Automatica*, **44**(8), 1958–1969 (2008).

- J. Rawlston, J. Guo, F. J. Schork, and M. A. Grover, "A kinetic Monte Carlo study on the nucleation mechanisms of oil-soluble initiators in the miniemulsion polymerization of styrene," *Journal of Polymer Science Part A Polymer Chemistry*, **46**(18) 6114–6128 (2008).

- A. Hernandez Moreno and M. Grover Gallivan, "An exploratory study of discrete time state-space models using kriging," *Proceedings of the 2008 American Control Conference* 3993–3998 (2008).

- C. Oguz, M. A. Gallivan, S. Cakir, E. Yilgor, and I. Yilgor, "Influence of polymerization procedure on polymer topology and other structural properties in highly branched polymers obtains by A_2+B_3 approach," *Polymer*, **49**, 1414–1424 (2008).

- A. F. Hernandez and M. A. Grover, "Stochastic dynamic predictions using kriging for nanoparticle synthesis," Proceedings of the 10th International Symposium of Process Systems Engineering, Bahia, Brazil, August 2009.

- J. Rawlston, F. J. Schork, and M. A. Grover, "Multiscale modeling of branch length in butyl acrylate," *Macromolecular Theory and Simulations*, accepted for publication (2010).

- A. F. Hernandez and M. A. Grover, "Stochastic dynamic predictions using Gaussian process models for nanoparticle synthesis," invited paper submitted to *Computers & Chemical Engineering* (2010).

Several additional publications on this research are expected, including two currently in preparation by Hernandez and Grover and one currently in preparation by Rawlston and Grover.

3 Synergistic Interactions

3.1 United Technologies

Professor Grover has fosted interactions and collaboration with United Technologies, which were initiated via the Dynamics and Control program at AFOSR. Through conversations with Dr. Andrzej Banaszuk at AFOSR program reviews, Professor Grover began discussions with UTRC in the area of empirical modeling of complex systems. Prof. Grover visited UTRC in 2007 and gave an invited seminar, meeting with researchers in controls and in materials. This ultimately led to collaboration with UTCPower in materials for fuel cells. Professor Grover first consulted with UTCPower, and then was awarded a research contract to optimize the process inputs for a nanoparticle synthesis process, using dynamic modeling and experiments. Consequently, this process became the primary case study used now for the GPM study under this grant. She visited UTCPower and UTRC again in late 2008, giving another invited seminar at UTRC and a progress report at UTCPower.

3.2 Air Force Research Laboratory

Professor Grover was awarded a United States Air Force Summer Faculty Fellowship in 2006, which supported her work with the Materials and Sensors Directorates at Wright Patterson Air Force Base in Dayton, Ohio (POC Dr. Donald Dorsey). The work focused on understanding the degradation dynamics of high power transistors made from compound semiconductors, and included modeling and directing of experiments. The nanoscale layers in the device exhibit unique electronic phenomenon that cannot be described using continuum models. This work led to continued interaction and discussions with AFRL, and also with Georgia Tech professor Samuel Graham, who performs transistor degradation experiments. Prof. Graham subsequently was also awarded a USAF Summer Faculty Fellowship, and then used it to also work with Dr. Dorsey at WPAFB.

4 Gaussian process modeling

The work on GPM is summarized here, since the key publications on this aspect of the research are not yet available in the archival literature.

4.1 Case study

As motivated by our collaborations with United Technologies, we apply the GPM approach for empirical modeling to the dynamics of nanoparticle synthesis, specifically the deposition of platinum nanoparticles on carbon nanotubes in a supercritical carbon dioxide process. The goal is to predict the time-evolution of the nanoparticle size distribution, as well as the efficiency of the reaction and the total amount of platinum deposited.

4.1.1 Precursor adsorption on nanotube

There are two stages in the deposition of the nanoparticles. First, the platinum precursor molecule must be adsorbed onto the surface of the carbon nanotubes, after which the process inputs are changed to induce a chemical reaction, releasing the elemental platinum so that it can self-organize into nanoparticles on the carbon nanotube surface. Key experimental data used to build the model are shown in Figure 1, along with model fits to this data.

If the platinum precursor is not soluble in the carbon dioxide, then the platinum cannot reach the carbon nanotube surface or react efficiently. Thus, the solubility of the precursor must be both

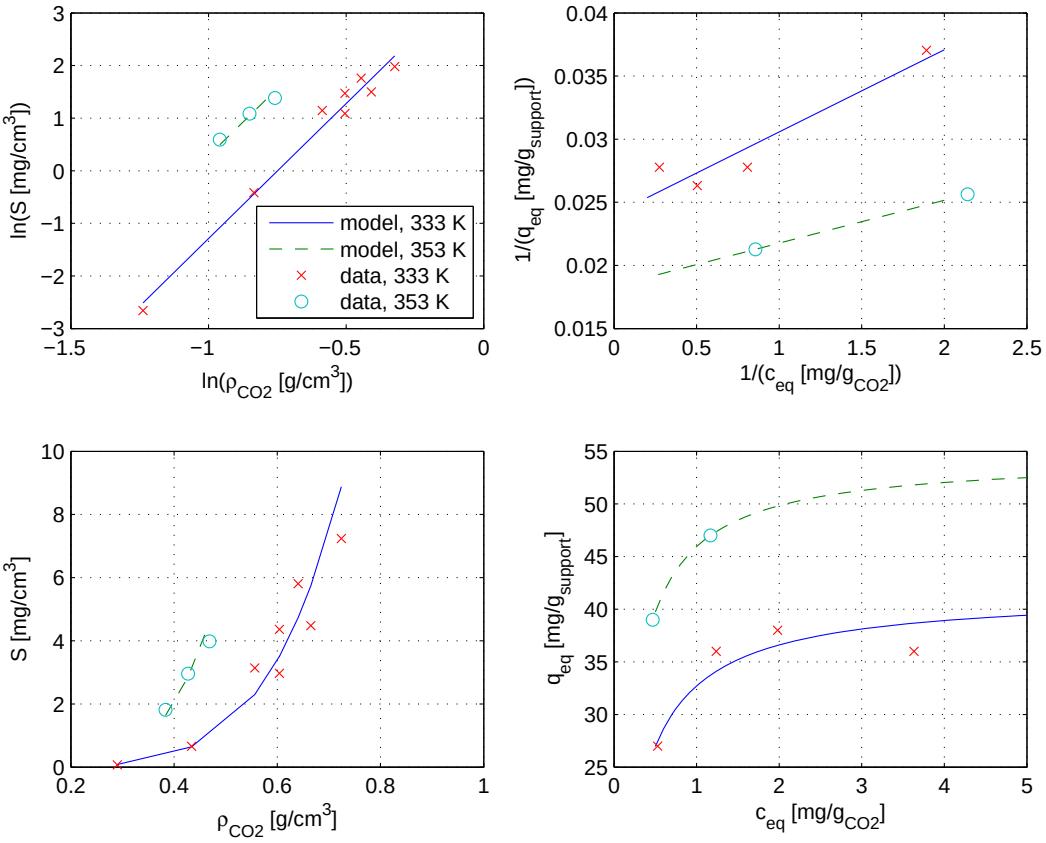


Figure 1: (left) Parameter fit for precursor solubility: k = 5.1, α = -9200 1/K, β = 32; (right) Parameter fit for absorption isotherm: at 333 K, $K_1 = 3.7$ g_{CO2}/mg_{prec}, $Q_0 = 42$ mg_{prec}/g_{sup}; at 353 K, $K_1 = 35.4$ g_{CO2}/mg_{prec}, $Q_0 = 54$ mg_{prec}/g_{sup}.

measured [2, 3] and modeled [4]. The Chrastil equation is used here to model the solubility S:

$$\ln S = k \ln \rho_{CO_2} + \frac{a}{T} + b \tag{3}$$

$$\rho_{CO_2} = f_{EOS}(T, P) \tag{4}$$

The density of carbon dioxide comes from an equation of state (EOS) [5]. The units on S are (mg $PtMe_2COD$)/(cm^3 CO_2).

Once the solubility of the precursor mass has been verified, the adsorption of the precursor on the carbon nanotube surface (CNT) must be predicted. An adsorption isotherm is used to quantify the adsorption capacity of the CNT:

$$q_{eq} = \frac{K_1 Q_0 c_{eq}}{1 + K_1 c_{eq}} \tag{5}$$

using experimental measurements [3] to estimate K_1 and Q_0. The amount of platinum precursor on the CNT is q_{eq}, while the amount in the fluid phase is c_{eq}.

The temperature dependence of the parameter K_1 modeled by

$$K_1(T) = K_0 \exp\left(\frac{-E_a}{RT}\right) \tag{6}$$

with E_a=1.8×10^4 J/mol and K_0=3.0×10^3 g_{CO2}/mg_{prec}.

4.1.2 Nanoparticle formation on nanotube

The formation of the platinum nanoparticles can be modeled with a birth-death process for $C_i(t)$ [6], where C_i is the concentration of nanoparticles with i atoms. The resulting population balance model is

$$\frac{dC_i}{dt} = G_{i-1}C_{i-1} - G_i C_i \tag{7}$$

$$C(0, t) = 0 \tag{8}$$

$$C(t, 0) = R_{nuc}(t) \tag{9}$$

The kinetic rate constants used here are estimated from the experimental studies of Bayrakceken and Erkey [7], to yield a mean particle diameter of 2 nm and a standard deviation of 0.5 nm. The growth of the nanoparticles is modeled via a kinetic Monte Carlo stochastic simulation, according to the stochastic simulation algorithm of Gillespie [8].

4.2 Key equations

Gaussian process modeling (GPM) is an empirical modeling approach for generalized linear regression models which formulates a local correlation between the residuals of the linear regression model as a function of the location of the model inputs [9, 10]. This concept is not usually employed in system identification and model reduction, in which the location of the experimental points is not explicitly considered once the model has been identified. Frequently, the local correlation between model inputs is described by the distance between them in a monotonically decaying function. Other correlation functions includes dot products and linear terms that emphasize correlation over specific locations in the input space and not just by the relative distance between points. Some examples of correlation functions can be found in [1, 11]

Assume there is a set $\mathcal{D}$ of n input/output pairs $\{\mathbf{X}_i, Y_i\}$, where $\mathbf{X}_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$, $i = 1 \dots n$. A usual distance-based covariance function to model the correlation between samples in the set is:

$$V_{ij}\left(\mathbf{X}_i, \mathbf{X}_j\right) = \sigma_c^2 \cdot \exp\left[-\frac{1}{2}\sum_{k=1}^{d}\frac{\left(X_{i,k} - X_{j,k}\right)^2}{\ell_k^2}\right] + \sigma_u^2 \cdot \delta_{ij} \tag{10}$$

where δ_{ij} is the Kronecker delta and $\boldsymbol{\theta} = \left[\sigma_c^2, \sigma_u^2, \ell_1^2 \dots \ell_d^2\right]$ are the kriging parameters that control the features of the correlation. In particular, σ_u^2 is included to model the random noise component of the function at each sample point.

By using this covariance function, a kriging prediction for a new input $\mathbf{x}_0$ is expressed under the best linear unbiased estimator as [12]:

$$[y|\mathbf{x}, \mathcal{D}] \sim \mathcal{N}\left(\hat{y}\left(\mathbf{x}, \mathcal{D}\right), \Sigma_y\left(\mathbf{x}, \mathcal{D}\right)\right) \tag{11}$$

$$\mathbb{E}\left[y|\mathbf{x}, \mathcal{D}\right] = \hat{y}\left(\mathbf{x}, \mathcal{D}\right) = \mathbf{h}^T\left(\mathbf{x}\right)\hat{\boldsymbol{\beta}} + \mathbf{v}^T(\mathbf{x}, \mathcal{D})V^{-1}(\mathcal{D}, \mathcal{D})\left[\mathbf{Y} - H(\mathcal{D})\hat{\boldsymbol{\beta}}\right] \tag{12}$$

$$\Sigma_y\left(\mathbf{x}, \mathcal{D}\right) = V\left(\mathbf{x}, \mathbf{x}\right) - \begin{bmatrix}\mathbf{h}^T\left(\mathbf{x}\right) & \mathbf{v}^T\left(\mathbf{x}, \mathcal{D}\right)\end{bmatrix}\begin{bmatrix}0 & H^T\left(\mathcal{D}\right) \\ H\left(\mathcal{D}\right) & V\left(\mathcal{D}, \mathcal{D}\right)\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{h}\left(\mathbf{x}\right) \\ \mathbf{v}\left(\mathbf{x}, \mathcal{D}\right)\end{bmatrix} \tag{13}$$

where $\mathbf{v}$ is the correlation vector between $\mathbf{x}$ and the inputs in $\mathcal{D}$ using Eq.(10), $\mathbf{h}, H$ represents a set of p regression functions evaluated at the unknown input and the inputs in $\mathcal{D}$ respectively, $\hat{\boldsymbol{\beta}}\left(\boldsymbol{\theta}\right) = \left(H^T V^{-1} H\right)^{-1} H^T V^{-1}\mathbf{Y}$ is the generalized least-squares estimator of the regression coefficients and $\mathbf{Y} = \left[Y_1, Y_2, \dots Y_n\right]^T$.

A key feature of our GPM research is the use of a GPM in a recursive dynamic formulation, which necessitates the consideration of uncertainty in the "input" $x[k]$ to the function F, as in Equation (1). Specifically, the general function input $\mathbf{x}$ has the following Gaussian distribution:

$$\mathbf{x} \sim \mathcal{N}\left(\hat{\mathbf{x}}, \Sigma_{\mathbf{x}}\right), \quad \hat{\mathbf{x}} \in \mathbb{R}^d, \Sigma_{\mathbf{x}} \in \mathbb{R}^{d \times d}$$

This input uncertainty is accounted for by making a Taylor series approximation around the input mean distribution $\hat{\mathbf{x}}$ and expanding Equations (12) and (13). Since uncertainty is not considered in the training set $\mathcal{D}$, this symbol will be drop out from now on. For simplicity in the notation and following derivation, the following notation is used.

Definition:

$$G \in \mathbb{R}^{p+n \times p+n} = \begin{bmatrix}0 & H^T \\ H & V\end{bmatrix}$$

$$\mathbf{g}\left(\mathbf{x}\right) \in \mathbb{R}^{p+n} = \begin{bmatrix}\mathbf{h}\left(\mathbf{x}\right) \\ \mathbf{v}\left(\mathbf{x}\right)\end{bmatrix}$$

Using G and $\mathbf{g}\left(\mathbf{x}\right)$, Equations (12) and (13) become

$$\hat{y}\left(\mathbf{x}\right) = \begin{bmatrix}\mathbf{0} & \mathbf{Y}^T\end{bmatrix} G^{-1}\mathbf{g}\left(\mathbf{x}\right) \tag{14}$$

$$\Sigma_y\left(\mathbf{x}\right) = V\left(\mathbf{x}, \mathbf{x}\right) - \mathbf{g}^T\left(\mathbf{x}\right) G^{-1}\mathbf{g}\left(\mathbf{x}\right) \tag{15}$$

The application of the Taylor-series approximation to Equations (14) and (15) leads to the following corrected expressions

$$\hat{y}\left(\mathbf{x}\right) \approx \hat{y}\left(\hat{\mathbf{x}}\right) + \left(\mathbf{x} - \hat{\mathbf{x}}\right)^T \left.\frac{\partial\left[\hat{y}\left(\mathbf{x}\right)\right]}{\partial x_i}\right|_{\mathbf{x}=\hat{\mathbf{x}}} + \frac{1}{2}\left(\mathbf{x} - \hat{\mathbf{x}}\right)^T \left.\frac{\partial^2\left[\hat{y}\left(\mathbf{x}\right)\right]}{\partial x_i \partial x_j}\right|_{\mathbf{x}=\hat{\mathbf{x}}} \left(\mathbf{x} - \hat{\mathbf{x}}\right)$$

$$\Sigma_y(\mathbf{x}) \approx \Sigma_y(\hat{\mathbf{x}}) + (\mathbf{x}-\hat{\mathbf{x}})^T \left.\frac{\partial[\Sigma_y(\mathbf{x})]}{\partial x_i}\right|_{\mathbf{x}=\hat{\mathbf{x}}} + \frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^T \left.\frac{\partial^2[\Sigma_y(\mathbf{x})]}{\partial x_i \partial x_j}\right|_{\mathbf{x}=\hat{\mathbf{x}}} (\mathbf{x}-\hat{\mathbf{x}})$$

$$\mathbb{E}[\hat{y}(\mathbf{x})] \approx \hat{y}(\hat{\mathbf{x}}) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} Cov(x_i, x_j) \left.\frac{\partial^2[\hat{y}(\mathbf{x})]}{\partial x_i \partial x_j}\right|_{\mathbf{x}=\hat{\mathbf{x}}} \tag{16}$$

$$\mathbb{E}[\Sigma_y(\mathbf{x})] \approx \Sigma_y(\hat{\mathbf{x}}) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} Cov(x_i, x_j) \left.\frac{\partial^2[\Sigma_y(\mathbf{x})]}{\partial x_i \partial x_j}\right|_{\mathbf{x}=\hat{\mathbf{x}}} \tag{17}$$

From Equations (14) and (15), matrix differential calculus is used to estimate the corresponding derivatives.

$$\frac{\partial^2[\hat{y}(\mathbf{x})]}{\partial x_i \partial x_j} = \begin{bmatrix}\mathbf{0} & \mathbf{Y}^T\end{bmatrix} G^{-1}\left[\frac{\partial^2[\mathbf{g}(\mathbf{x})]}{\partial x_i \partial x_j}\right] \tag{18}$$

$$\begin{aligned}\frac{\partial^2[\Sigma_y(\mathbf{x})]}{\partial x_i \partial x_j} = &-\left[\frac{\partial^2[\mathbf{g}(\mathbf{x})]}{\partial x_i \partial x_j}\right]^T G^{-1}\mathbf{g}(\mathbf{x}) - \left[\frac{\partial[\mathbf{g}(\mathbf{x})]}{\partial x_j}\right]^T G^{-1}\left[\frac{\partial[\mathbf{g}(\mathbf{x})]}{\partial x_i}\right] \\ &-\left[\frac{\partial[\mathbf{g}(\mathbf{x})]}{\partial x_i}\right]^T G^{-1}\left[\frac{\partial[\mathbf{g}(\mathbf{x})]}{\partial x_j}\right] - \mathbf{g}^T(\mathbf{x}) G^{-1}\left[\frac{\partial^2[\mathbf{g}(\mathbf{x})]}{\partial x_i \partial x_j}\right]\end{aligned} \tag{19}$$

The Taylor-series approximation depends on the derivatives of the function $\mathbf{g}(\mathbf{x})$. The definition of derivatives for vectors in matrix differential calculus describes the derivative to each element in the vector (different from the Jacobian vector or Hessian matrix, where the elements in the vectors and matrices are based on the derivatives of a scalar function). Then, by definition

$$\frac{\partial[\mathbf{g}(\mathbf{x})]}{\partial x_i} \in \mathbb{R}^{p+n}, \quad \frac{\partial[\mathbf{g}(\mathbf{x})]}{\partial x_i} = \left[\frac{\partial[\mathbf{h}(\mathbf{x})]}{\partial x_i} \quad \frac{\partial[\mathbf{v}(\mathbf{x})]}{\partial x_i}\right] \tag{20}$$

where

$$\frac{\partial[\mathbf{h}(\mathbf{x})]}{\partial x_i} \in \mathbb{R}^{p}, \quad \frac{\partial[\mathbf{h}(\mathbf{x})]}{\partial x_i} = \left[\frac{\partial[h_1(\mathbf{x})]}{\partial x_i}, \frac{\partial[h_2(\mathbf{x})]}{\partial x_i}, \cdots, \frac{\partial[h_p(\mathbf{x})]}{\partial x_i}\right]$$

is the first derivative of each of the p regression functions used in the GPM model respect to a single x_i entry in the input vector. This implies that each h_i regression function must be explict on the elements of the input vector $\mathbf{x}$. Similarly

$$\frac{\partial[\mathbf{v}(\mathbf{x})]}{\partial x_i} \in \mathbb{R}^{n}, \quad \frac{\partial[\mathbf{v}(\mathbf{x})]}{\partial x_i} = \left[\frac{\partial[v_1(\mathbf{x})]}{\partial x_i}, \frac{\partial[v_2(\mathbf{x})]}{\partial x_i}, \cdots, \frac{\partial[v_n(\mathbf{x})]}{\partial x_i}\right]$$

where v_k is the spatial correlation between the input vector $\mathbf{x}$ and a query point $\mathbf{X}_k$ in the training data $\mathcal{D}$, using a correlation function. In particular, Equation (10) is often employed to represent the spatial correlation. This correlation function is known as Gaussian correlation function.

$$v_k(\mathbf{x}) = V(\mathbf{x}, \mathbf{X}_k) = \sigma_c^2 \cdot \exp\left[-\frac{1}{2}\sum_{k=1}^{d}\frac{(x_i - X_{k,i})^2}{\ell_i^2}\right] + \sigma_u^2 \cdot \delta_k$$

Using the same arguments presented previously, we can see that:

$$\frac{\partial^2[\mathbf{g}(\mathbf{x})]}{\partial x_i \partial x_j} \in \mathbb{R}^{p+n} \quad \frac{\partial^2[\mathbf{g}(\mathbf{x})]}{\partial x_i \partial x_j} = \left[\frac{\partial^2[\mathbf{h}(\mathbf{x})]}{\partial x_i \partial x_j} \quad \frac{\partial^2[\mathbf{v}(\mathbf{x})]}{\partial x_i \partial x_j}\right] \tag{21}$$

To complete the description of these derivatives, the first and second derivative expressions for the Gaussian correlation function are given:

$$\frac{\partial\left[v_k\left(\mathbf{x}\right)\right]}{\partial x_i} = V\left(\mathbf{x}, \mathbf{X}_k\right)\left[-\frac{\left(x_i - X_{k,i}\right)}{\ell_i^2}\right] \tag{22}$$

$$\frac{\partial^2\left[v_k\left(\mathbf{x}\right)\right]}{\partial x_i \partial x_j} = \begin{cases} V\left(\mathbf{x}, \mathbf{X}_k\right)\left[-\frac{\left(x_i - X_{k,i}\right)}{\ell_i^2}\right]^2 + V\left(\mathbf{x}, \mathbf{X}_k\right)\left[-\frac{1}{\ell_i^2}\right] & \text{if } i = j \\ V\left(\mathbf{x}, \mathbf{X}_k\right)\left[-\frac{\left(x_i - X_{k,i}\right)}{\ell_i^2}\right]\left[-\frac{\left(x_j - X_{k,j}\right)}{\ell_j^2}\right] & \text{if } i \neq j \end{cases} \tag{23}$$

4.3 Comparison of sampling strategies

The application of GPM in dynamic systems modeling can be viewed as the identification of a mapping function of the state variables from one time to the next one, which is applied recursively to predict dynamic trajectories. This modeling framework assumes that the one-step-ahead dynamics can be represented using the Markov property, making the prediction dependent of the location of the state variables in the state space. We refer to the one-step prediction as the "local prediction." Because the modeling framework describes the dynamic trajectory as a series of local predictions at each discrete time step, a sequence of predictions is used to evaluate the overall or global dynamic performance of the approximated model. We refer to this type of prediction as the "global prediction."

Depending on the different settings used to create the GPM, the accuracy of the approximated model will change, both locally and globally. Table 1 summarizes the combinations of settings that we used to construct the comparison of sampling strategies. A total of 168 potential combinations have been evaluated. Because of the random component in the selection of sample points and the nature of the kMC simulation as the source of information for model building, a single realization of a particular combination of settings in not enough to capture the dynamics for the approximated model. Each of the potential combinations has been repeated 10 times and the local and global performance of the approximated model have been averaged over those realizations.

Settings	**Options**
Number of sample points	30, 45, 60, 75, 90, 105, 120
Repetition level	1, 3, 5
Sampling strategy	1. From pre-collected points in dynamic trajectories 2. Random walk approach in convex hull
Regression Functions	1. Constant (Ordinary kriging, OK) 2. Forward stepwise regression, linear functions (Universal Kriging, UK)
Parameter Estimation	Maximum likelihood estimation (MLE) or restricted MLE

Table 1: Summary of potential combinations to create approximated models, evaluated for local and global prediction analysis.

4.3.1 Local Prediction Analysis

Figure 2 represents the graphical description of the local prediction analysis. A total of 100 input points have been randomly selected over the relevant dynamic region defined by the convex hull of the kMC simulations. The random selection over the multidimensional state space was made by

a random walk approach, evaluating the inequality constraints that defines the region. For each of the input points, a reconstructed version of the kMC state variables has been created. Using the kMC information, a one-step-ahead kMC simulation has been replicated 5 times to create a sample mean vector $\bar{\mathbf{x}}(k+1)$ and sample variance vector $S^2_{\mathbf{x}}(k+1)$ that summarizes the repetition information over the five state variables. These vectors are compared against the vectors of expected mean prediction, Eq.(11), and prediction variance, Eq.(13), from each of the kriging models. The Euclidean distance between the vectors is used as a local error metric to quantify the accuracy of the approximated model. The local error metric (LEM) for the expected mean prediction is:

$$LEM = \frac{1}{10 \cdot 100} \sum_{r=1}^{10} \sum_{j=1}^{100} \left[\sqrt[2]{\frac{1}{d} \sum_{i=1}^{d} \left(\bar{x}_i (k+1) [\mathbf{x}_j(k)] - \mathbb{E} [\hat{x}_i (k+1) | \mathbf{x}_j(k), \mathcal{D}_r] \right)^2} \right] \tag{24}$$

where the index r indicates different realizations of the same combination of settings to build the approximated model. A similar expression can be derived for the sample variance and prediction variance vectors.

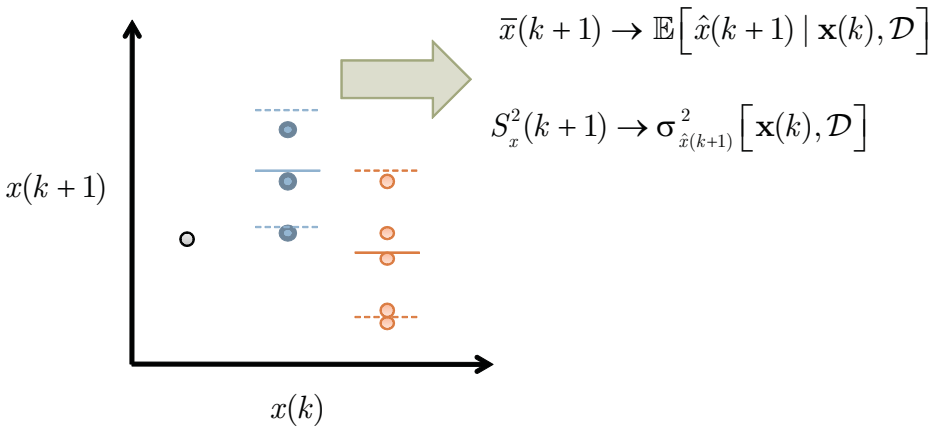
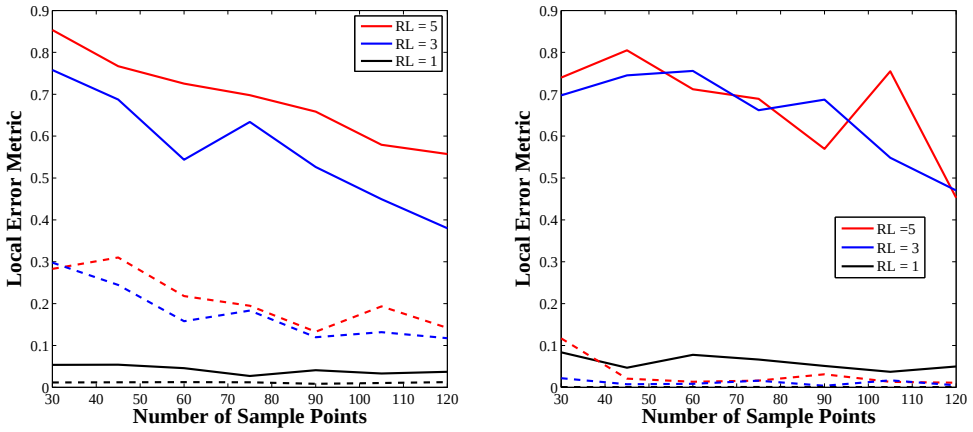


Figure 2: Graphical description of local prediction analysis. The plot shows a one-dimensional case of the local prediction analysis but it can easily interpreted for multidimensional space.

The evaluation of the local error metric for different settings of the approximated model and sampling strategies are presented in Figures 3 and 4. The presence of global regression functions in the approximated model improves one-step-ahead prediction for all repetition levels and number of sample points compare to the constant regression function. Once should expect to see a decrease in the local error metric as the number of sample points increases, and this is confirmed by the figures. Similar conclusions and numerical values can be obtained with both MLE and RMLE parameter estimation methods.

Given that kriging is a distance-based empirical model, the spatial coverage of the input points over the dynamic region is important for a good prediction. When a repetition level of 1 is used to create the approximated model (i.e. no repetitions), there is a better coverage over the dynamic region compared to the other repetition levels because the number of input points in the convex hull is higher. This situation could explain why the local error metric is lower at low repetition levels. This spatial coverage issue could be used also to explain the differences between the sampling strategies. Since the information from pre-collected dynamic trajectories is used to create the convex hull, it is likely that those points are at the boundaries of the convex hull, offering a poor

performance in the interior of the region. On the other hand, the random walk could offer a better coverage of the points because the procedure is exploratory over the constrained region.



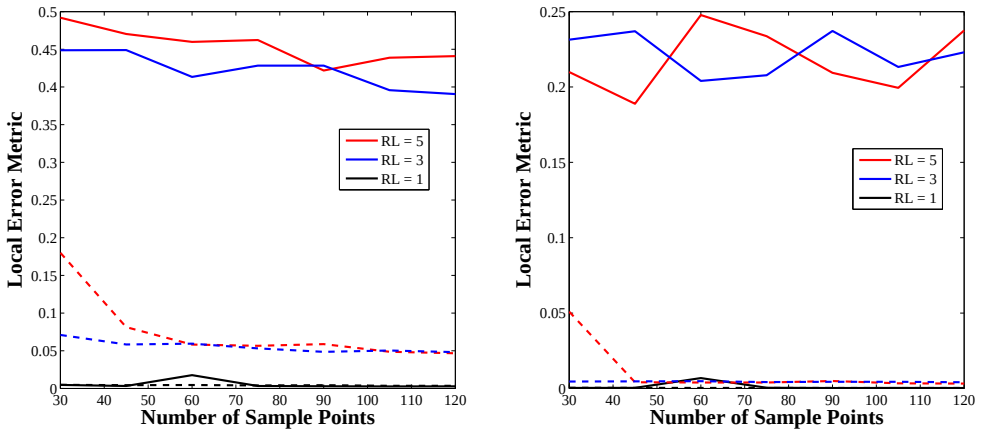
(a) (b)

Figure 3: Local prediction analysis for different repetition levels using points from pre-collected dynamic trajectories. Parameter estimator: MLE. (a) Expected mean vector. (b) Prediction variance vector. — Ordinary Kriging, - - - Universal Kriging. RL = Repetition Level.

Surprisingly, the use of repetitions in the model does not improve the local prediction variance. This is explained by the poor spatial coverage that a higher repetition level implies (remember the trade-off between repetition level and number of input points in the convex hull) or by the assumption that the noise component of the state variables is constant over the entire dynamic region, according to the covariance function that is used in the approximated model, see Eq.(14). Other covariance functions can be implemented to enhance the identification of the noise component in the kMC simulation.

4.3.2 Global Prediction Analysis

Figure 5 is a graphical representation of the global prediction analysis. Global prediction analysis is performed to evaluate the prediction capability of the approximated model over the operating space of precursor mass and carbon nanotube mass at the beginning of the process. Different dynamic trajectories are obtained for different initial state variables from the operating space. A total of 40 different points in the operating space have been randomly selected using a Latin hypercube design, to represent the potential initial conditions of the system. For a single combination of precursor mass and carbon nanotube mass, 5 repetitions of the kMC dynamic trajectory are simulated. Since all dynamic trajectories are collected with the same sampling time, at each discrete time point in the trajectories a sample mean vector and a sample variance vector can be computed for all state variables, describing a time series of means and variances for that particular operating point. This procedure can be repeated for each of the 40 selected operating points and compared with the expected mean vector and prediction variances described by the predicted dynamic trajectory for a particular operating point. Based on this comparison a dynamic error metric (DEM) for the



(a) (b)

Figure 4: Local prediction analysis for different repetition levels using random points selected in the convex hull. Parameter estimator: MLE. (a) Expected mean vector. (b) Prediction variance vector. — Ordinary Kriging, - - - Universal Kriging. RL = Repetition Level.

expected mean prediction can be defined as

$$DEM = \frac{1}{10 \cdot 40 \cdot k_T} \sum_{r=1}^{10} \sum_{j=1}^{40} \sum_{k=0}^{k_T} \left[\sqrt{\frac{1}{d} \sum_{i=1}^{d} \left(\bar{x}_{i,j}\left(k+1\right)\left[\mathbf{x}_j(k)\right] - \mathbb{E}\left[\hat{x}_{i,j}\left(k+1\right)|\mathbb{E}\left[\hat{\mathbf{x}}_j(k)\right], \mathcal{D}_r\right]\right)^2} \right] \tag{25}$$

where the index j describes different dynamic trajectories and r refers to different realizations of the same combination of settings in the approximated model. Notice that in this analysis, the evaluation of the expected mean prediction uses the prediction from the previous discrete time point.

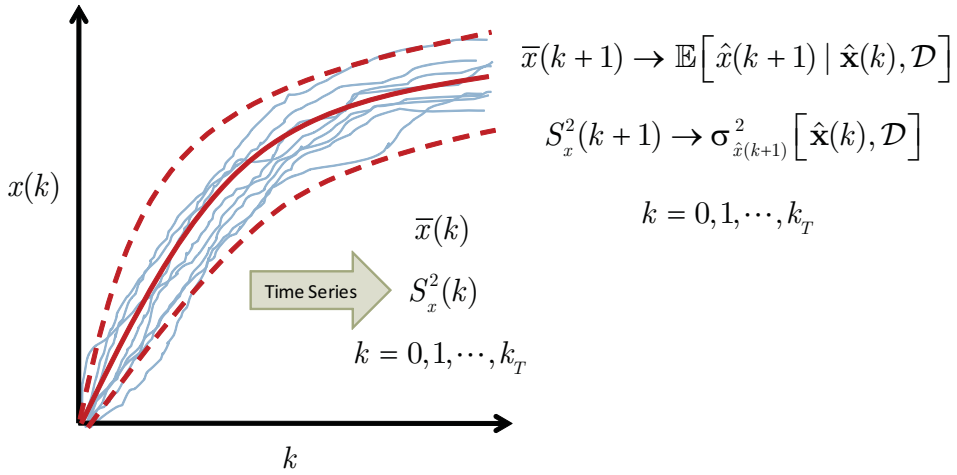
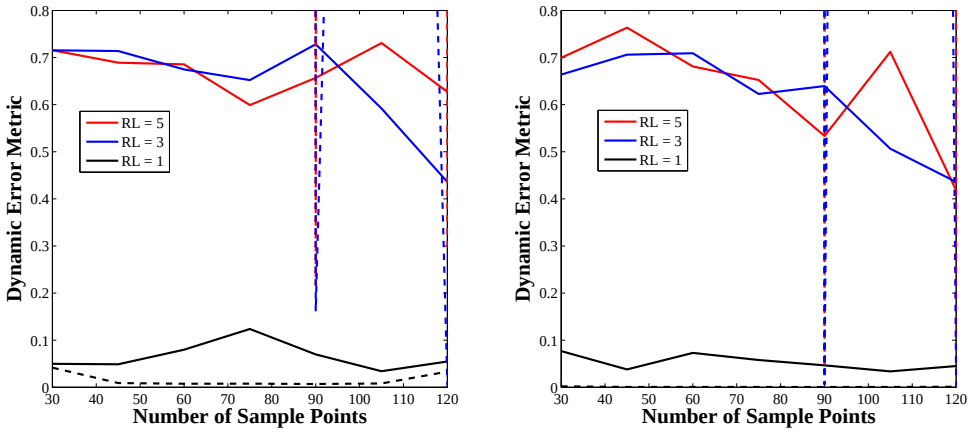


Figure 5: Graphical description of global prediction analysis. The plot shows a one-dimensional case of the local prediction analysis but it can easily interpreted for multidimensional space.

Figures 6 and 7 describe the conclusions of the global prediction analysis. In principle, if the local prediction is accurate over the dynamic region, then we should have a good global prediction as well. The plots indicate that the uses of linear regression functions makes poor global predictions, compared with constant regression functions. An possible explanation for this situation is the lack of information to create a compact minimal dynamic region. Because of the use of these linear regression functions, the predictions might be extrapolating outside the dynamic region that we initially defined. Although the addition of the trend functions via $\mathbf{h}(\mathbf{x})$ intuitively seems to be useful for improving dynamics predictions, in our implementation we find that using ordinary kriging in fact provides more robust prediction by avoiding extrapolation.



(a) (b)

Figure 6: Global prediction analysis for different repetition levels using points from pre-collected dynamic trajectories. Parameter estimator: MLE. (a) Expected mean vector. (b) Prediction variance vector. — Ordinary Kriging, - - - Universal Kriging. RL = Repetition Level.

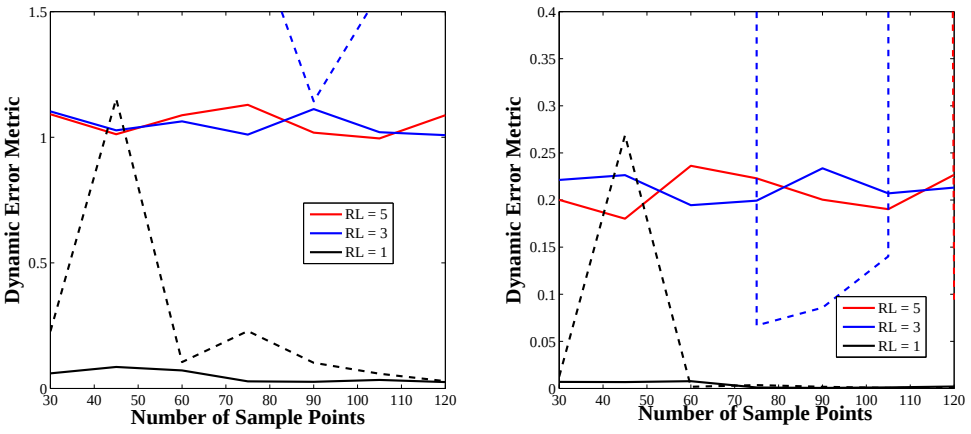
References

[1] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[2] O. Aschenbrenner, S. Kemper, N. Dahmen, K. Schaber, and E. Dinjus. Solubility of β-diketonates, cyclopentadienyls, and cyclooctadiene complexes with various metals in supercritical carbon dioxide. *Journal of Supercritical Fluids*, 41(2):179–186, 2007.

[3] O. Aschenbrenner, N. Dahmen, K. Schaber, and E. Dinjus. Adsorption of dimethyl(1,5-cyclooctadiene)platinum on porous supports in supercritical carbon dioxide. *Industrial & Engineering Chemistry Research*, 47(9):3150–3155, 2008.

[4] S. Yoda, Y. Mizuno, T. Furuya, K. Otake, T. Tsuji, and T. Hiaki. Solubility measurements of noble metal acetylacetonates in supercritical carbon dioxide by high performance liquid chromatography (HPLC). *Journal of Supercritical Fluids*, 44(2):139–147, 2008.



(a) (b)

Figure 7: Global prediction analysis for different repetition levels using random points selected in the convex hull. Parameter estimator: MLE. (a) Expected mean vector. (b) Prediction variance vector. — Ordinary Kriging, - - - Universal Kriging. RL = Repetition Level.

[5] R. Span and W. Wagner. A new equation of state for carbon dioxide covering the fluid region from the triple-point temperature to 1100 K at pressures up to 800 MPa. *Journal of Physical and Chemical Reference Data*, 29(6):1361–1433, 2000.

[6] E. E. Finney and R. G. Finke. Nanocluster nucleation and growth kinetic and mechanistic studies: A review emphasizing transition-metal nanoclusters. *Journal of Colloid and Interface Science*, 317(2):351–374, 2008.

[7] A. Bayrakceken, U. Kitkamthorn, M. Aindow, and C. Erkey. Decoration of multi-wall carbon nanotubes with platinum nanoparticles using supercritical deposition with thermodynamic control of metal loading. *Scripta Materialia*, 56(2):101–103, 2007.

[8] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434, 1976.

[9] Noel Cressie. *Statistics for Spatial Data*. Wiley Interscience, 3 edition, 1993.

[10] J. R. Koehler and A. B. Owen. Computer experiments. In S. Ghosh and C. R. Rao, editors, *Handbook of Statistics*, pages 261–308. Elsevier Science, New York, 1996.

[11] S. N. Lophaven, H. B. Nielsen, and J. Sondergaard. *DACE: A Matlab Kriging toolbox, v. 2.0*. IMM Technical University of Denmark, Lyngby, 2002.

[12] Arthur S. Goldberger. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57(298):369–375, 1962.